SAIC-94/1155

REV-10/95

Global Association

Final Report

R. Le Bras, H. Swanger, T. Sereno, G. Beall, R. Jenkins

W. Nagy, and A. Henson

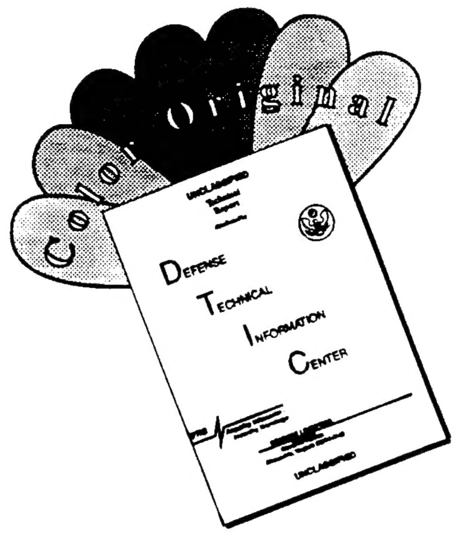
November 23, 1994

Science Applications International Corporation

10260 Campus Point Drive

San Diego, California 92121

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF COLOR PAGES WHICH DO NOT REPRODUCE LEGIBLY ON BLACK AND WHITE MICROFICHE.

SAIC-94/1155

REV-10/95

Global Association

Final Report

R. Le Bras, H. Swanger, T. Sereno, G. Beall, R. Jenkins
W. Nagy, and A. Henson

November 23, 1994

Science Applications International Corporation
10260 Campus Point Drive
San Diego, California 92121

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
23 November 1994

3. REPORT TYPE AND DATES COVERED
Technical Report May 1994 - Nov 1994

4. TITLE AND SUBTITLE
Global Association Final Report

5. FUNDING NUMBERS
F08606-90-D-0005

6. AUTHORS
R. Le Bras, H. Swanger, T. Sereno, G. Beall, R. Jenkins, W. Nagy, and A. Henson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Science Applications International Corporation
10260 Campus Point Drive
San Diego, California 92121

8. PERFORMING ORGANIZATION REPORT NUMBER
SAIC-94/1155

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
HQ Air Force Technical Applications Center (HQ AFTAC/TTR)
1030 S. Highway A1A
Patrick AFB FL 32925-3002

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This Final Report describes the *Global Association System* for automatic interpretation of seismic data to associate signals and locate seismic events. It includes results for a synthetic data set intended to represent the type of global network that is planned for the upcoming Group of Scientific Experts Third Technical Test called GSETT-3 (for an overview, see *Kerr* [1993]), and actual data from the prototype U.S. National Data Center (PNDC) at AFTAC. This system uses a hybrid method that integrates techniques in generalized beam forming [e.g., *Ringdal and Kværna*, 1989; *Taylor and Leonard*, 1992; *Leonard*, 1993] and expert systems [*Bache et. el.*, 1993]. It is designed to handle the large volumes of data that are needed to monitor compliance with a Comprehensive Test Ban Treaty (CTBT).

14. SUBJECT TERMS

Global Association	GA	ESAL
Location	Event Formation	SAIC-94/1155

15. NUMBER OF PAGES
29

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFIC ATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	Same as report

NSN 7540-01-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std. Z39-18

(This page intentionally left blank.)

Table of Contents

List of Figures

Page

List of Tables

Page

1.0 Introduction

This Final Report describes the *Global Association System* for automatic interpretation of seismic data to associate signals and locate seismic events. It includes results for a synthetic data set intended to represent the type of global network that is planned for the upcoming Group of Scientific Experts Third Technical Test called GSETT-3 (for an overview, see *Kerr* [1993]), and actual data from the prototype U.S. National Data Center (PNDC) at AFTAC. This system uses a hybrid method that integrates techniques in generalized beam forming [e.g., *Ringdal and Kværna*, 1989; *Taylor and Leonard*, 1992; *Leonard*, 1993] and expert systems [*Bache et. el.*, 1993]. It is designed to handle the large volumes of data that are needed to monitor compliance with a Comprehensive Test Ban Treaty (CTBT).

1.1 Background

In 1993, discussions began concerning the design of the International Data Center (IDC) and U.S. National Data Center (NDC) for the upcoming GSETT-3 experiment. This experiment will be the first demonstration of a global monitoring system that addresses the CTBT problem. In the early discussions, the global network was envisioned to include as many as 60 primary stations (mostly arrays) to provide continuous data, and up to 200 secondary stations to provide waveform segments upon request. The volume of data (~10 Gbytes per day) and number of events (300-400 per day) were estimated to be approximately a factor of five to ten times greater than encountered from existing global networks.

SAIC performed an engineering study to assess whether existing seismic monitoring systems could be modified to handle these expected data volumes. The software components of the ADSN (*AFTAC Distributed Subsurface Network*) and IMS (*Intelligent Monitoring System*) were considered. The conclusion of this study was that the primary bottleneck would be the automatic association and location program, **ESAL** [*Bratt et al.*, 1991, 1994]. Performance analyses indicated that **ESAL**'s execution time scaled roughly with the square of the detection density. This was unacceptable since extrapolation to the estimated data volumes for a CTBT monitoring network indicated that **ESAL** would not be able to process the data in real time.

To address this concern, SAIC proposed to replace **ESAL** with a new hybrid method. This method splits the tasks currently performed by **ESAL** into separate modules that can be run in parallel (Figure 1). The main association module is very similar to published generalized beam forming techniques and to the unpublished technique used by the Australian IDC in the GSETT-2 experiment [Ken Muirhead, personal communication]. The main difference is that a model of the probability of detection is used to significantly reduce the search space. Initial work on the new hybrid method was supported by the Advanced Research Projects Agency (ARPA). That work was continued under this AFTAC task order.

1.2 Summary of Accomplishments

The major accomplishments of this task order include:

Initial release of the Global Association System: This system has been developed and installed on the PNDC at AFTAC. All modules have been built and tested using Sun Workstations under the Solaris 2.3 Operating System. The system includes three new

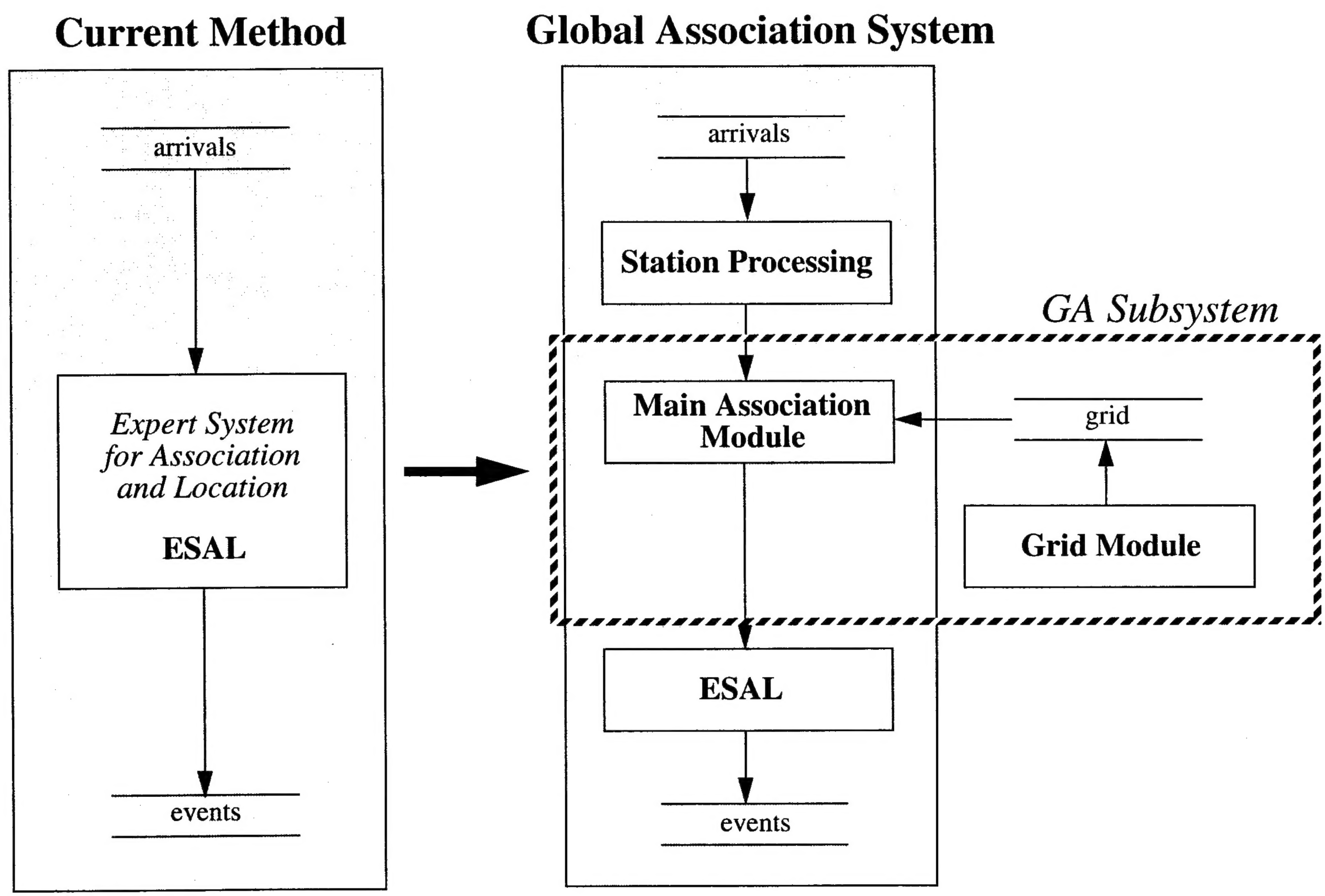


Figure 1. The current and planned configuration of the software for automated association and location are shown.

modules: **StaPro** performs station processing; **GAcons** constructs the generic knowledge base on a global grid; and **GAassoc** generates preliminary event hypotheses. Also, **ESAL** was enhanced to resolve conflicts among the preliminary event hypotheses formed by **GAassoc**.

Port to Oracle 7: All new modules use the Generic Database Interface (GDI) to support both Version 6 and 7 of the Oracle database [*Anderson et al.*, 1994].

Design Document and User's Manual: Documentation describing the system, software components and algorithms was delivered under this Task Order [*Le Bras et al.*, 1994].

1.3 Report Outline

The emphasis of this Final Report is on test results. The *Global Association System* and the algorithms that it uses are described in detail in our Design Document and User's Manual [*Le Bras et al.*, 1994]. Section 2 summarizes this design, and Section 3 presents the test results. Section 3.1 gives results for station processing (**StaPro**). Section 3.2 describes the synthetic data set that we used to represent the type of global network that is planned for the GSETT-3 experiment. Sections 3.3 and 3.4 describe the results of applying the *Global Association System* to the synthetic data set. Results on computational efficiency and the quality of the seismic bulletin are reported. Section 3.5 compares the performance of the *Global Association System* to the performance of **ESAL**. Section 3.6 gives preliminary results from tests at the PNDC. Finally, Section 4.0 summarizes our recommendations for future development and testing.

2.0 System Description

This section summarizes the *Global Association System.* The first subsection summarizes the major components of the system and the relationship between them. The second subsection provides a brief description of the algorithms used by each of the major components. These algorithms are described in detail in our Design Document and User's Manual [*Le Bras et al.*, 1994]. The last subsection is an illustrative example of event formation through various stages in the processing sequence.

2.1 Overview

The main components of the *Global Association System* are:

StaPro: This module performs station processing. It analyzes detections and their features to make preliminary seismic phase identifications. It also forms single-station location and magnitude hypotheses that are used by **GAassoc** to screen detections from small events.

GAcons: This module builds one or several global grid files containing the knowledge base for the association process. Overlapping circular grid cells provide complete global coverage, including depth cells in areas where deep seismicity is known to occur.

GAassoc: This module identifies event hypotheses using an exhaustive search over all grid cells. It uses the information in the grid file produced by **GAcons** to identify detections that are consistent with a particular event hypothesis.

EServer/ESAL: These modules resolve conflicts (i.e., phases that are associated to more than one event) and refine the event hypotheses formed by **GAassoc**. **EServer** provides the interface which moves data between the external file format used by **ESAL** and the relational database management system (RDBMS).

The high-level processing and data flow for the *Global Association System* are shown in Figure 2. **GAcons** is not included because it is not part of the real-time processing. It produces a static grid file that must be recomputed only if the network changes or modifications are made to the knowledge base. The Task Controller could be an automated or manual process. In the simplest configuration, it is a user that manually initiates each component after the previous one has completed. **StaPro** is initiated by the Task Controller after completion of signal detection and feature extraction for a particular station. **GAassoc** performs the global association after **StaPro** has completed for all stations. Information about detections is transferred from the individual **StaPro** runs to **GAassoc** through the RDBMS. **GAassoc** writes all event hypotheses and associations to the RDBMS. **EServer** reads these hypotheses, prepares external input files, and initiates **ESAL** to resolve conflicts, refine event hypotheses, and produce the final bulletin. Finally, **EServer** writes **ESAL**'s results to the RDBMS.

The *Global Association System* was developed on Sun workstations under the Solaris 2.3 Operating System. The current version uses an Oracle 7.1.3 RDBMS. The new components (**StaPro, GAcons**, and **GAassoc**) use the Generic Database Interface (GDI) for all database transactions [*Anderson et al.*, 1994]. **StaPro** uses CLIPS Version 6.0 to provide run-time configurability of station-specific rules. CLIPS is a knowledge-based macro language which is supported by NASA. **ESAL** is programmed in the ART (Automated Reasoning Tool) expert system shell from Inference Corporation.

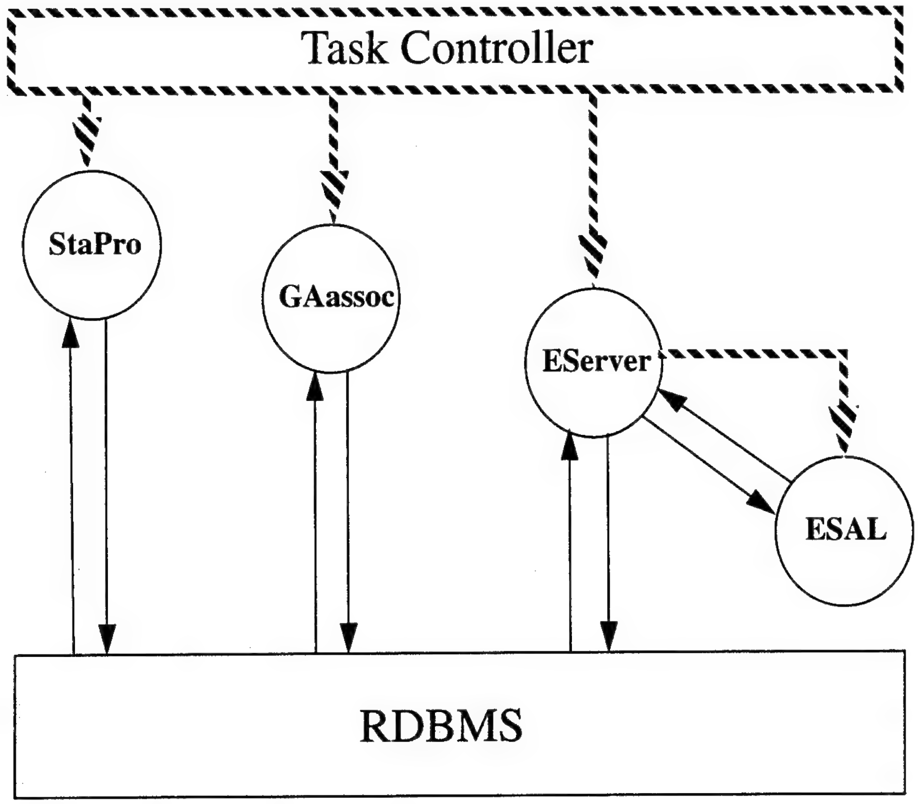


Figure 2. This shows the high-level processing and data flow for the *Global Association System*. Process flow is indicated by dashed lines, and data flow is indicated by solid lines.

2.2 Algorithms

The main algorithms used by the *Global Association System* are described briefly in this section. Detailed descriptions can be found in our Design Document [*Le Bras et al.*, 1994].

2.2.1 Station Processing (StaPro)

Station processing (**StaPro**) determines the initial wave type of each detection (Teleseism, Regional *P*, Regional *S*, or Noise), groups detections that appear to come from the same event, and determines a preliminary identification of the seismic phase (*P*, *Tx*, *Pn*, *Pg*, *Px*, *Sn*, *Lg*, *Rg*, *Sx*, or *N*). **StaPro** contains the same logic as **ESAL** for this task, and it includes the ability to compute single-station location and magnitude hypotheses [*Bratt et al.*, 1991, 1994].

The initial wave-type identification is based on a combination of slowness from *f-k* analysis, polarization attributes, and frequency content. The specific combination depends on the station type (i.e., array or three-component), and whether a neural network or rules are used [*Bratt et al.*, 1991,1994; *Sereno and Patnaik*, 1993].

After all detections have been assigned an initial wave type, they are collected into groups that appear to come from the same event. The first arrival in each group is called the *generator*, and

its initial wave type must be either *P* or *T.* Subsequent arrivals are added to the group on the basis of their compatibility (based on azimuth and amplitude) with the *generator.* Three categories of events are treated separately in grouping: teleseismic, local and regional. Teleseismic and local groups are found first, and their detections are removed from consideration for regional grouping. All regional *P*-wave detections that were not previously assigned to a local group are potential *generators* for a regional group [*Bratt et al.*, 1991,1994].

The third major task in station processing is preliminary seismic phase identification. The current implementation uses simple rules to identify teleseismic and local phases, and a more sophisticated approach based on Bayesian analysis to identify regional phases. The regional *generator* is identified first, and either the first or the largest *S* wave in the group is identified by Bayesian analysis. The Bayesian analysis technique uses conditional probabilities based on slowness from *f-k* analysis, horizontal-to-vertical power ratio, *S-P* time, period, and context with respect to other *S* phases. The remaining phases in a regional group are identified by prediction [*Bratt et al.*, 1991,1994; *Bache et al.*, 1993].

The final task is to compute a single-station event location and estimate local magnitude for association groups that satisfy user-specified event confirmation criteria. Event confirmation is based on a weighted-count of defining observations (arrival time, azimuth, and slowness). Local magnitudes are estimated for confirmed events using the method described by *Bache et al.* [1991] in their Appendix A. The local magnitude is used by **GAassoc** to screen detections from small events.

2.2.2 Gridding (GAcons)

GAcons builds a global grid of precomputed information for use in **GAassoc**. The first task of **GAcons** is to establish a quasi-uniformly distributed set of points on a sphere. The points represent the centers of circular surface cells providing a complete overlapping coverage of the Earth. The grid is completed by adding depth cells with their center at the same latitude and longitude as the surface cells and their depth at user-specified values. The depth cells are added in areas where deep seismicity is known to occur. The gridding algorithms and the format of the grid file are described in detail in our Design Document [*Le Bras et al.*, 1994].

For each cell, **GAcons** establishes a list of stations that have a non-negligible probability of detecting the earliest arrival for an event within the cell. The probability level is specified by the user. Events are simulated for each grid cell. Their locations are distributed uniformly within the cell volume and their magnitudes are distributed according to a user-specified recurrence rate (i.e., *b*-value). The probability of detecting these events at each station in the network is estimated from attenuation models and noise estimates. For each simulated event, the station that records the earliest arrival is added to the list of "first-arrival stations." This list is used by **GAassoc** to significantly reduce the search space required for global association.

2.2.3 Event Formation (GAassoc and ESAL)

The *Global Association System* combines gridded search (**GAassoc**) and expert system (**ESAL**) techniques to associate signals and locate seismic events. Figure 3 shows the major tasks that are performed by each module. **GAassoc** uses the grid data generated by **GAcons** to associate arrivals processed by **StaPro** to form event hypotheses. Each instance of **GAassoc** executes a loop over all grid cells included in the input grid file. Grid files can be generated for multiple

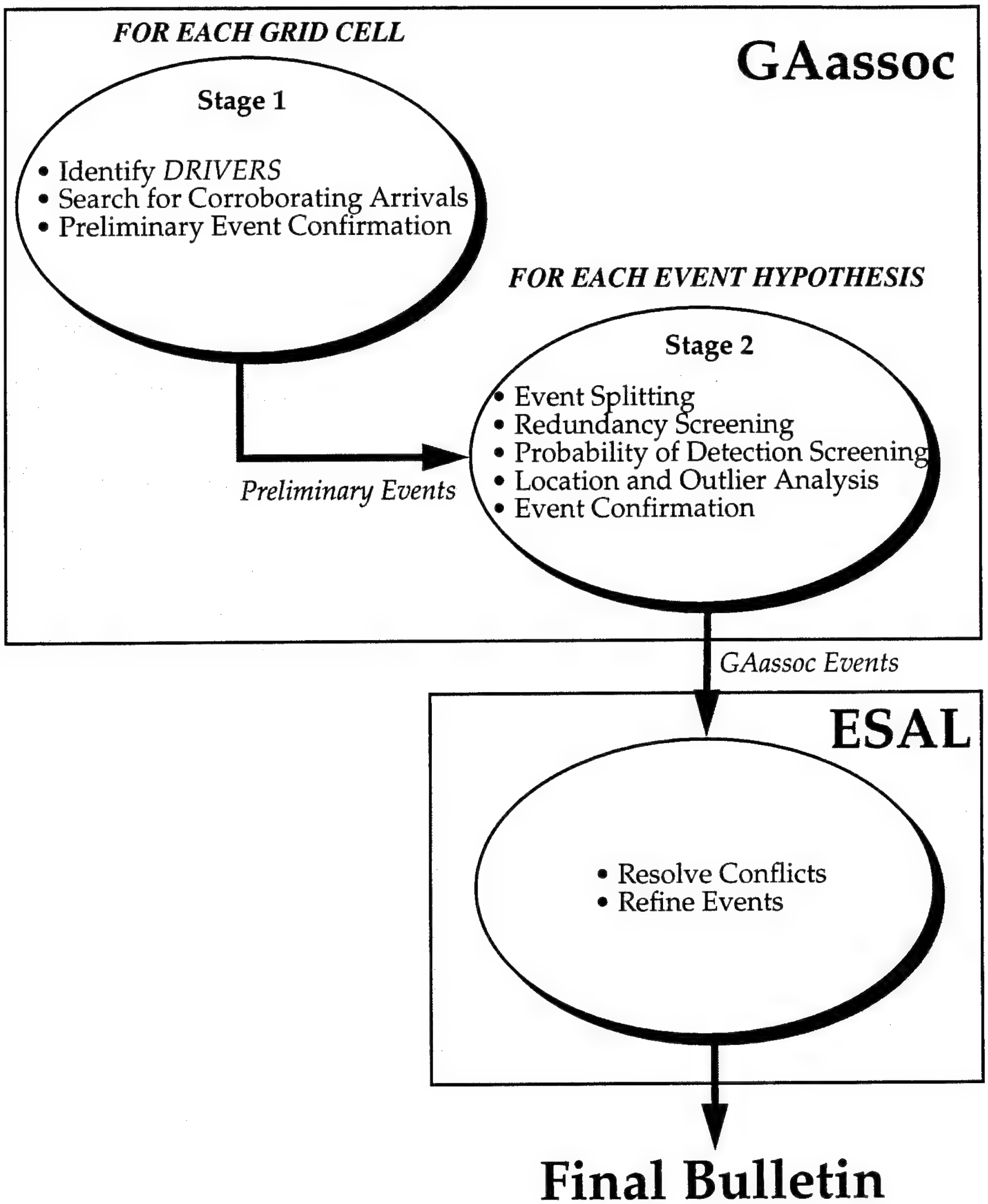


Figure 3. This shows the major steps in the event formation process. **GAassoc** forms a preliminary event list by examining each grid cell as a potential event location. It then performs several tasks on each preliminary event to form a condensed list of events for **ESAL** to process (*GAassoc Events*). **ESAL** resolves conflicts and refines the event hypotheses to produce a final bulletin.

sectors, and **GAassoc** can be run separately (in parallel) for each sector.

GAassoc has been divided into two stages. The first stage examines each grid cell as a potential seismic event location. It performs three tasks:

GAassoc (Stage 1)

- *Identify DRIVERS:* A *DRIVER* is an arrival at one of a limited set of stations in the network that could record the earliest arrival for an event in the given grid cell. The *DRIVER* has to have been identified as *P*, *Pn* or *Pg* by station processing (**StaPro**) and it must satisfy constraints on its slowness vector.

- *Search for Corroborating Arrivals:* A simple screening process based on travel time and slowness vector (if available) is used to search for arrivals that are consistent with the *DRIVER*. A more rigorous chi-square statistical test is applied if an arrival successfully passes this initial screening.

- *Preliminary Event Confirmation:* Association groups are eliminated if they do not satisfy an event confirmation test based on a weighted-count of the number of defining observations (arrival time, azimuth, and slowness).

This results in a set of preliminary event hypotheses for each grid cell. After all grid cells in the current sector have been processed, the second stage of **GAassoc** performs the following additional tasks for each event hypothesis:

GAassoc (Stage 2)

- *Event Splitting:* The preliminary events may include incompatible arrivals such as two or more arrivals at the same station identified as the same phase, or the same arrival identified as two or more different phases. When this occurs, the degeneracy is split into two or more separate, self-consistent events.

- *Redundancy Screening:* The same set of associations (or a subset of them) can be consistent with two adjacent grid points. The redundant hypotheses are removed from the preliminary event list.

- *Probability of Detection Screening:* A network probability test can be applied to remove hypotheses from the preliminary event list that are formed by an unlikely combination of stations.

- *Location and Outlier Analysis:* Events remaining after the preliminary screening are located and an analysis is made of the residuals. Outliers are removed if necessary, redundancy checks performed, and the location is refined.

- *Event Confirmation:* A number of confirmation tests can be applied to each event. These include the weighted-count test described above, a supplemental restriction on the number of associated arrivals, a restriction on the size of the location error ellipse, and the network probability test that was applied before location (this time using the computed event location and magnitude).

After all sectors are completed, **ESAL** is applied to the results from **GAassoc** to resolve conflicts (i.e., phases that are associated to more than one event) and refine the event hypotheses.

The two major tasks are:

ESAL

- *Resolve Conflicts*: **ESAL** resolves conflicts among different **GAassoc** preliminary hypotheses. Each arrival is disassociated from all but the "best" event hypothesis. Four tests are currently available to determine which one is "best". Three of the tests are arrival-by-arrival tests, including smallest error ellipse, the largest number of defining phases, and a composite test that depends on an ordered list of criteria; one is an event-by-event test based on the largest number of defining phases. **ESAL** also resolves conflicts among events formed during the current time step and the preceding time step.

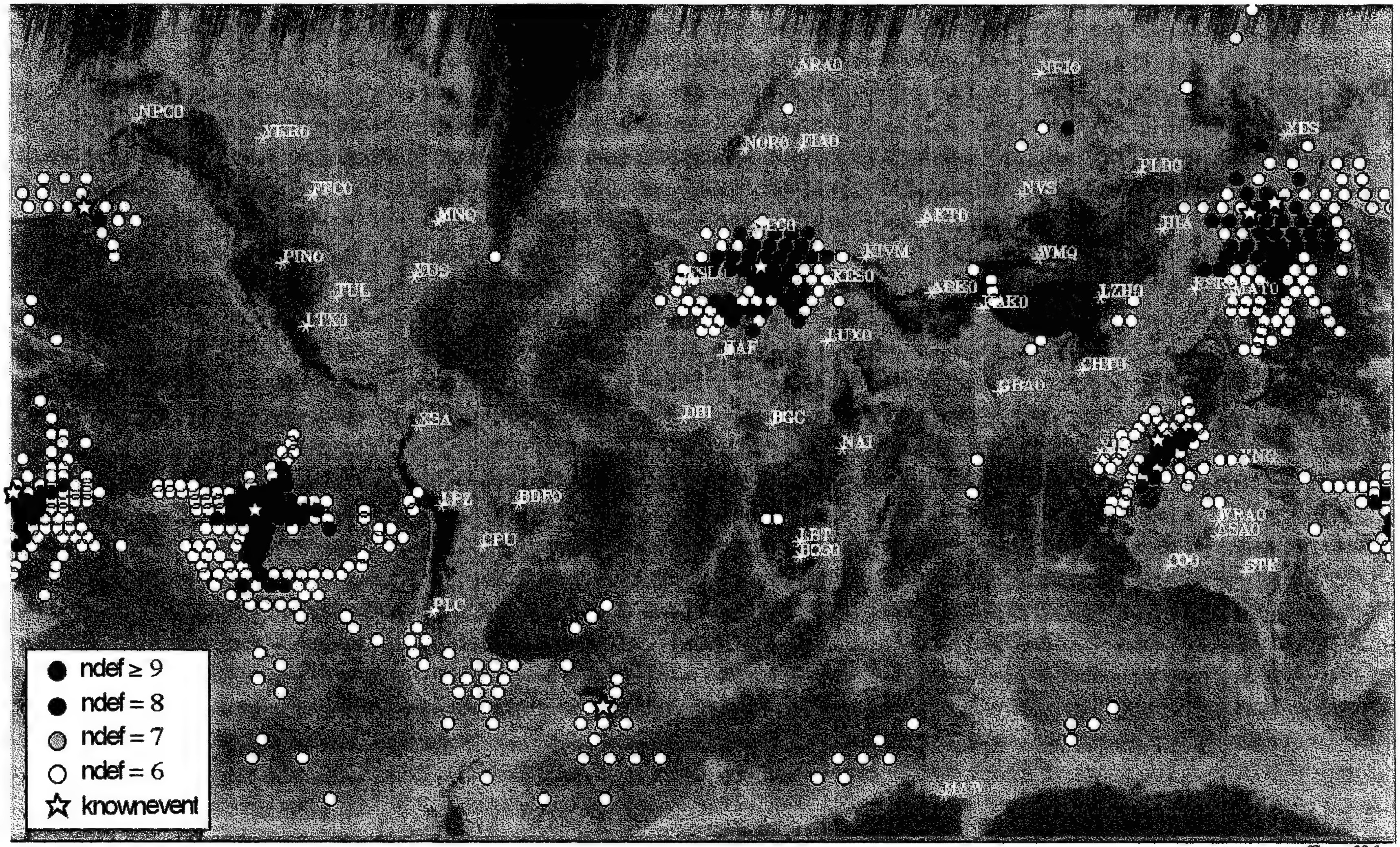
- *Refine Events:* **ESAL** associates late-arriving secondary phases and primary phases that may have been missed by **GAassoc**. It relocates the events and applies the same event confirmation tests that are applied by **GAassoc**.

The event formation process used by the *Global Association System* is described in more detail in our Design Document [*Le Bras et al.*, 1994].

2.3 Example

This section provides an illustrative example of the results from **GAassoc** at various stages in its processing sequence. We use the synthetic data set described in Section 3.2 to represent the upcoming GSETT-3 experiment. The hypothetical global network consists of 28 arrays and 24 three-component stations. The final data set includes an average of 314 events/day under normal (i.e., non-swarm) conditions.

GAassoc was applied to a 5-day synthetic data set using a grid spacing of approximately 3 degrees. The results of this test are described in Sections 3.3 and 3.4. In this section, we closely examine the results for a representative one-hour period. This period includes eight events with at least six defining phases. The event depths are between 5 and 500 km and the *mb* magnitudes are between 3.0 and 3.9. The preliminary events formed in the first stage of **GAassoc** processing are shown in Figure 4. The known locations of the hypothetical events are plotted as stars, and the hypotheses from **GAassoc** are color-coded by the number of defining phases. **GAassoc** forms clusters of event hypotheses for each of the known events, and very few clusters where there are no events. Most of these events are eliminated by the second stage of **GAassoc** processing. This is illustrated in Figure 5 which shows the events that remain after application of the second stage. On average, application of the second stage reduces the number of event hypotheses by about a factor of nine. All eight known events are still formed, and there are many fewer false alarms.



cfSereno28.fm

Figure 4. Preliminary events from the first stage of **GAassoc** are plotted for a one-hour segment of the synthetic GSETT-3 data. The preliminary events are color-coded by the number of defining observations. The location of the known events are plotted as stars.

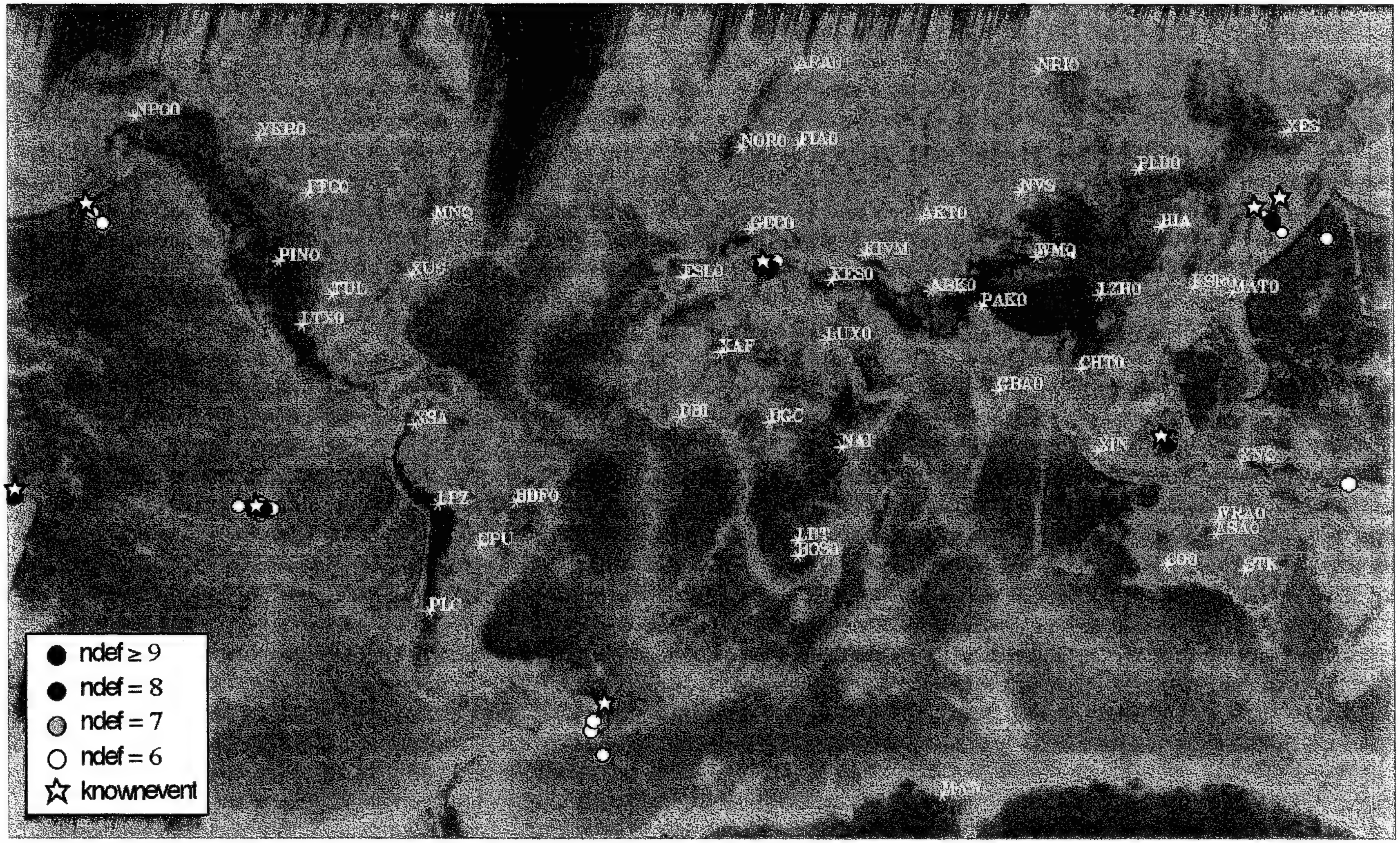


Figure 5. Events from the second stage of **GAassoc** are plotted for a one-hour segment of the synthetic GSETT-3 data. The preliminary events are color-coded by the number of defining observations. The location of the known events are plotted as stars.

3.0 Test Results

This section describes results of testing the *Global Association System*. **StaPro** test results are given in Section 3.1 for data recorded by the *Intelligent Monitoring System* (IMS). Other elements of the *Global Association System* were tested using two different data sets. The majority of the tests were conducted on a synthetic data set intended to represent the type of global network that is planned for the GSETT-3 experiment. This data set is described in Section 3.2. Test results are given for computational efficiency (Section 3.3) and the quality of the seismic bulletin that is produced (Section 3.4). In addition, Section 3.5 gives a direct comparison of the results from the *Global Association System* to those from **ESAL** using a representative subset of the synthetic data. The second test data set consists of real data from the Prototype National Data Center (PNDC) at AFTAC. These results are given in Section 3.6.

3.1 Station Processing (StaPro)

The primary motivation for porting **ESAL**'s station processing to a separate module is to improve the station characterization (parameters and rules can be customized for each station) and the processing speed. In the tests described below we demonstrate that **StaPro** produces the same results as **ESAL**'s station processing for three-component stations and arrays when given the same input. We also demonstrate that **StaPro** performs this task about three times faster than **ESAL**.

3.1.1 Array Station (ARA0) Unit Test

Data from the *Intelligent Monitoring System* (IMS) recorded by the ARCESS array in northern Norway (ARA0) were used to test **StaPro** for array stations. We processed data from a one-week continuous interval using both **StaPro** and **ESAL**. This interval included 3339 detections. The **StaPro** results were identical to the **ESAL** results. The comparison includes the initial wave types, phase groupings and initial phase identifications. This same test was repeated using one long time interval (as opposed to 20-minute segments) to verify that edge effects are handled correctly. The results were the same as they were for the segmented interval.

3.1.2 Three-Component Station (GAR) Unit Test

Data recorded by the IRIS station in Garm, Tajikistan (GAR) were used to test **StaPro** for three-component stations. We processed data from a one-week continuous interval using **StaPro** and **ESAL**. This interval included 1388 detections. We performed two tests: one using the neural network [*Sereno and Patnaik*, 1993] and one using the default rules for initial wave-type identification. **StaPro**'s results matched **ESAL**'s results identically for 98.6% of the detections when the neural network was used, and 92.1% when the default rules were used. All of the discrepancies were traced to a error in **ESAL**'s station processing regarding the revision of the initial wave type from a teleseism to a regional *P* if a compatible regional *S* is found [*Bratt et al.*, 1991, 1994]. This error has been fixed in **ESAL**.

3.1.3 Computational Efficiency

StaPro was found to run about three times faster than **ESAL**'s station processing during the tests described above. Table 1 compares the run time for **StaPro** and **ESAL**'s station processing

for 24-hour segments of array data (ARA0) and three-component data (GAR). **StaPro** and **ESAL** were run using 20-min processing intervals.

Table 1: StaPro and ESAL run times for 24-hour data segments

Station Type	ESAL	StaPro
Array Station (ARA0)	23 min	8 min
3-Component Station (GAR)	22 min	7 min

3.2 Synthetic Data Set

We developed a synthetic data set for the type of global network that is planned for the upcoming GSETT-3 experiment. The synthetic data were produced by the *Synthetic Detection Generator* (SDG) *Program Package* developed by R. North at the Geological Survey of Canada. The documentation for this program package is on-line at the Center for Monitoring Research (CMR) in Arlington, Virginia (formally the Center for Seismic Studies, or CSS).

We generated a list of hypothetical events for a five-day period using the **mkevents** program in the SDG *Program Package*. This program generates hypothetical events whose locations, depths, magnitudes and frequency of occurrence closely approximate the global distribution of natural seismicity. We specified the minimum log(seismic moment) of the events to be 19.5 which corresponds approximately to *mb* 2.8. We also used the **swarm** program to add an earthquake swarm centered at (43 N, 45 E) to the first day. The swarm activity rate, expressed as the proportion of the total global activity, was set to 3.0. This means that swarm events are produced at a rate equivalent to 3 times that for global seismicity.

The hypothetical GSETT-3 seismic network consists of 28 arrays and 24 three-component stations. The station locations are shown on the maps in Figures 4 and 5 (the arrays have 4-letter station names and the three-component stations have 3-letter station names). We used the values recommended by R. North in his SDG documentation for most of the station parameters. We assumed that the average noise level at each array is 0.6 nm (0-to-peak) at 1 Hz which includes noise suppression from beamforming. We assumed that the average noise level at each three-component station is 2.0 nm at 1 Hz. The number of events per day are summarized in Table 2, and the magnitude distribution is shown in Figure 6.

Table 2: Event Summary for GSETT-3 Synthetic Data Set

	DAY 1 (swarm)	DAY 2	DAY 3	DAY 4	DAY 5	TOTAL
NSTA ≥ 4	510	230	252	268	259	1519
NSTA = 3	71	31	36	37	37	212
NSTA = 2	53	25	28	26	28	160
TOTAL	634	286	316	331	324	1891

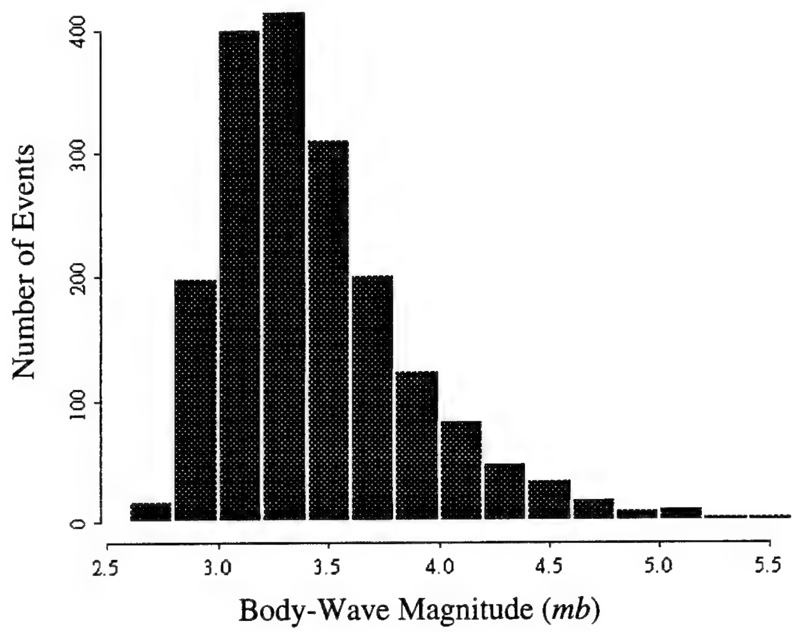
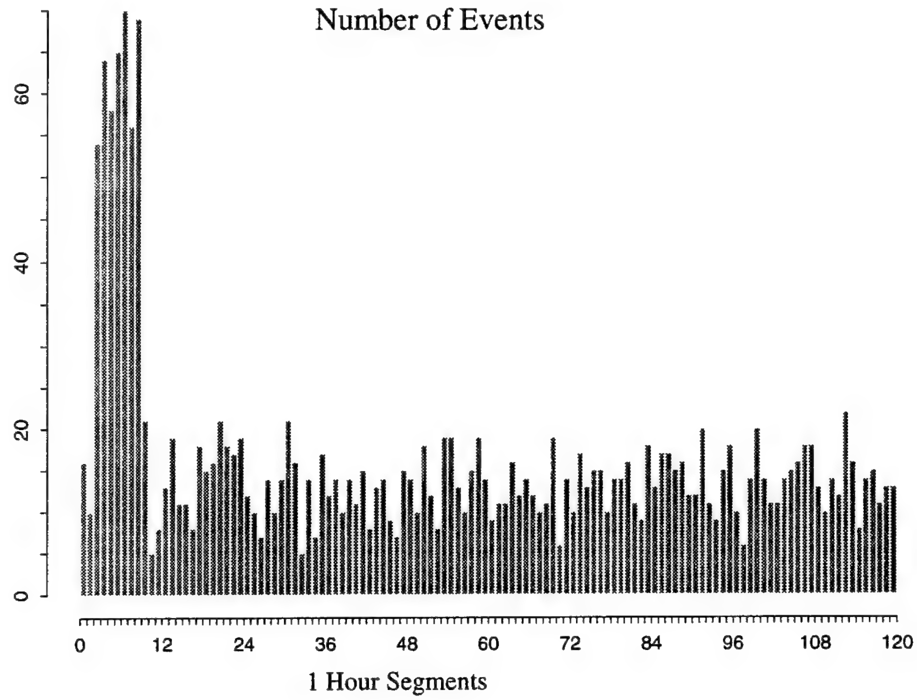


Figure 6. This shows the distribution of *mb* for the events in the synthetic GSETT-3 data set.

Synthetic detections were generated for the hypothetical network using the **arriv** and **ampmask** programs in the SDG *Program Package* based on the assumed station noise levels and an attenuation model for each phase. Random detections are also added to simulate false-alarms. The detection rates are summarized in Table 3. The average number of detections/station/day are listed separately for arrays and three-component stations. The two columns for each station type correspond to the swarm day (Day 1) and normal days (Days 2-5). The number of associated and unassociated detections are listed for each station type. The associated arrivals are also divided into separate counts for primary (*Pn*, *Pg*, *P*, *PKPab*, *PKPbc*, *PKPdf*) and secondary phases. The distribution of the events and detections in time are shown in Figure 7.

Table 3: Number of Detections/Station/Day in the Synthetic GSETT-3 data set

	ARRAYS Day 1	**ARRAYS Days 2-5**	**3-C Day 1**	**3-C Days 2-5**
Number of Detections	*418*	*268*	*215*	*199*
Associated	323	175	75	61
Primary	174	86	43	32
Secondary	149	89	32	29
Unassociated	95	93	140	138



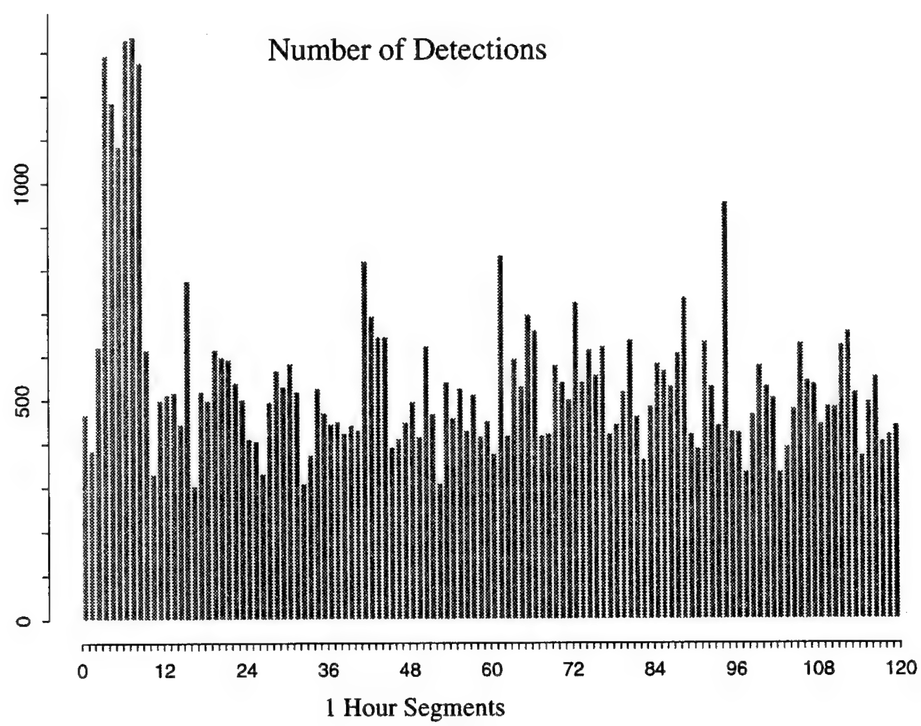


Figure 7. The number of events and number of detections are plotted for each one-hour interval in the synthetic data set. The swarm conditions on the first day are clearly visible.

The synthetic detection rates are generally consistent with rates observed by stations in the IDC ALPHA network (see the IDC Performance Reports which are available at the Center for Monitoring Research in Arlington, VA). However, the phase distributions are different. For example, the synthetic data set has a much lower percentage of regional and local phases than are reported by the IDC ALPHA stations.[1] Many of these phases are associated with small single-station events. The synthetic data set also has a higher percentage of teleseismic primary phases than are reported by the IDC ALPHA stations.[2] These factors contribute to detection rates for teleseismic primary phases which are significantly higher than those reported at the IDC ALPHA stations. Since each of these is a candidate for a *DRIVER* arrival in **GAassoc**, it is likely that this data set will take longer to process than data recorded by the actual global network that will be used in the GSETT-3 experiment. We believe therefore that our computational efficiency estimates in Section 3.3 are conservative.

The detection features used in the association process (e.g., time, azimuth, slowness) are generated from an assumed normal distribution. The variance of each attribute can depend on the signal-to-noise ratio and station type. We used variance estimates for each attribute based on our experience with IMS, IDC and ADSN data. For example, the synthetic azimuth residuals for arrays and three-component stations are plotted in Figure 8. Similar plots can be made for slowness, arrival time, rectilinearity and horizontal-to-vertical power ratios (the last two are used by **StaPro** for initial wave type identification for three-component stations).

3.3 Computational Efficiency

This section reports results on the computational efficiency of the *Global Association System* for the 5-day synthetic data set described in the previous section. These data were processed in a simulated pipeline using 20-min segments with a 20-min look back. The grid used for this test was for the whole Earth (parallelization was not used).

The cumulative CPU times for **GAassoc**, **ESAL** and **GAassoc+ESAL** are plotted in Figure 9 for the 5-day data set. The 20-min segments were combined into 120 segments, each one hour in length. The dark bars indicate the time spent in **GAassoc** and the white bars show the time spent in **ESAL**. The total height of each bar indicates the total CPU time needed to process that one-hour segment of data. The dashed horizontal line is drawn at one hour, so all segments below that line were processed in less than real time and all segments above it were not. With the exception of the swarm activity between 4 and 10 hours, most intervals could be processed in less than real time.

The processing time required by **GAassoc** depends primarily on the number of event hypotheses which must be considered and evaluated. This, in turn, depends on multiple external factors including the density of the detection data, the uncertainty in the detection features, and the number of associations for large events.[3] A comparison of the times in Figure 9 to the event and

1. Approximately 45% of the detections in the synthetic data set were determined to be local or regional by Station Processing compared to 60% of the non-noise detections at the IDC.
2. Approximately 90% of the teleseismic detections in the synthetic data set were determined to be primary phases by Station Processing compared to 75% at the IDC.
3. Large events will frequently form large clusters of preliminary event hypotheses in surrounding cells, as shown in Figure 4.

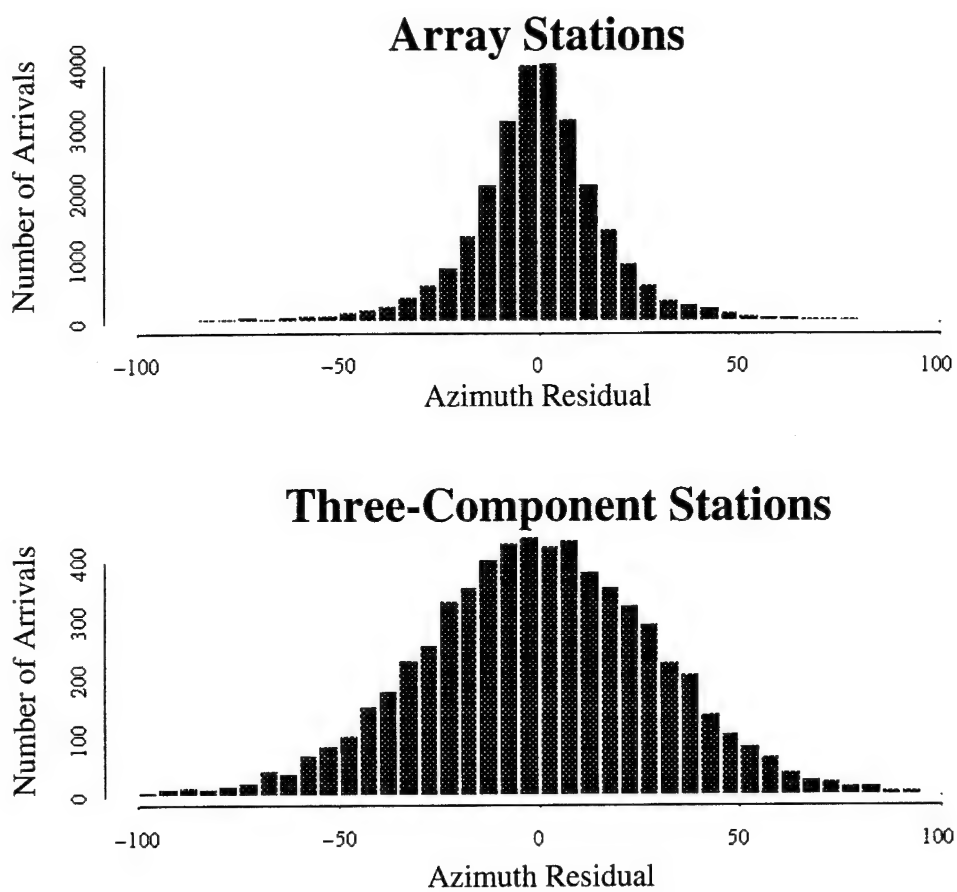


Figure 8. The distributions of the azimuth residuals in the synthetic data set are plotted for array stations (top) and three-component stations (bottom).

Figure 7 shows that, for this data set, the processing time for **GAassoc** is more dependent on the number of detections than the number of events. This is clearly the case for some of the peaks in the non-swarm periods.

ESAL processing times are proportionately higher within the swarm period than **GAassoc** processing times. The average ratio of **ESAL** CPU to **GAassoc** CPU is 2.78 during the swarm as compared to 1.62 during periods of normal seismicity. This suggests that the gridded approach may be more efficient for large volumes of swarm data than **ESAL**.

Figure 10 shows histograms of the CPU time to process 20-min segments for **GAassoc**, **ESAL** and **GAassoc+ESAL**. The data for the swarm period is separated from the periods of normal seismicity. During normal periods, the average CPU time for **GAassoc** is 2-4 min with an extreme value of 26 min, and the average CPU time for **ESAL** 4-6 min with an extreme value of 20 min. As noted above, the variance in CPU time is much higher for **ESAL** than for **GAassoc** during swarm conditions. Most of the variability in the CPU time for **GAassoc** is due to the time spent in the location module. The time in the main association module (Stage 1) is relatively constant.

The test results in Figures 9 and 10 are encouraging, but there are some periods that are not

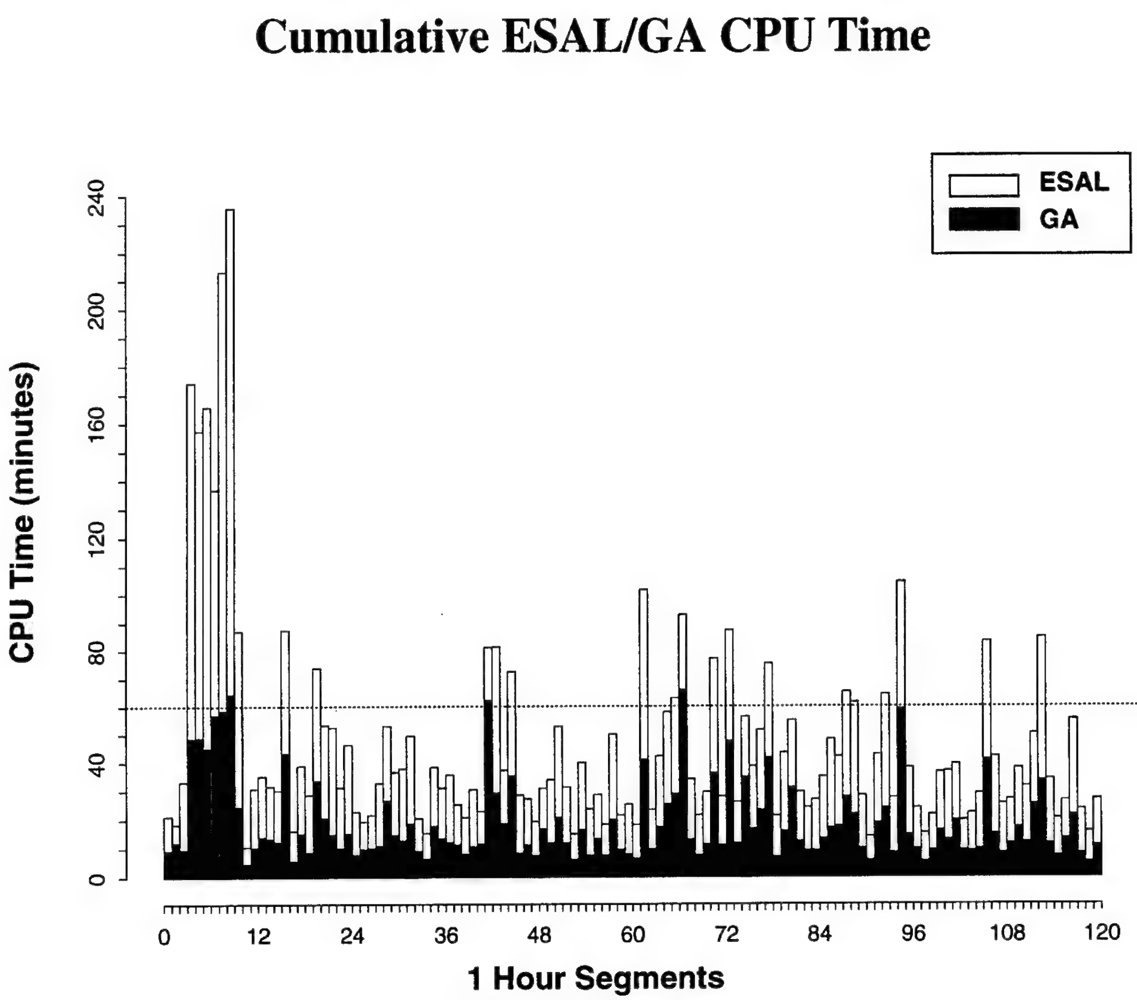


Figure 9. Cumulative CPU time is plotted for **GAassoc** plus **ESAL** for each one-hour time segment in the 5 days of synthetic data. The dark grey bars show the portion of time spent in **GAassoc**. The white bars shows the time spent in **ESAL**. Note the distinctly higher values in the first portion of day 1 (hours 4 to 10) corresponding to swarm activity.

processed in real time (even under normal conditions). However, these results are for a whole Earth grid. They do not take advantage of the ability of **GAassoc** to process different sectors in parallel. Also, much of **ESAL**'s time is spent on I/O which could be eliminated if conflict resolution and event refinement were added to **GAassoc** (see Section 3.5). Finally, we believe that the synthetic data set may have a higher density of potential *DRIVERS* than real data, so it could take longer to process. In summary, the *Global Association System* is very close to meeting the requirement of processing the large volumes of data from a CTBT monitoring network in real time under normal seismic activity. There is more work to be done to handle swarm conditions in real time, and we are working on this problem under separate ARPA sponsorship. We have several suggestions for improving the *Global Association System*, and these are described in more detail in Section 4.0.

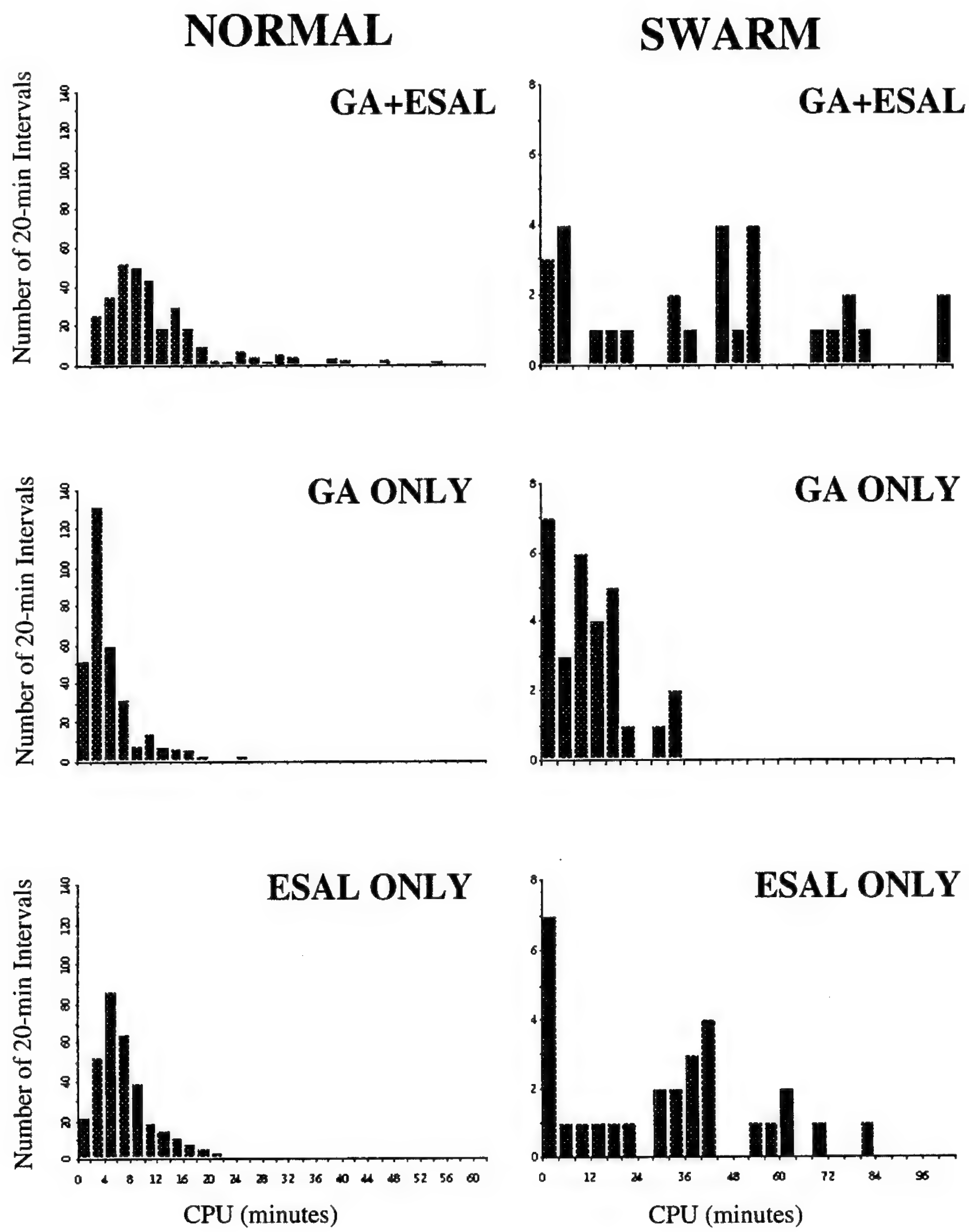


Figure 10. Histograms of CPU time spent in **ESAL, GA** and **GA+ESAL** for all 20-min segments in the 5-day synthetic data set. The three histograms to the left are for normal seismicity, and the three on the right are for the swarm period on the first day.

3.4 Bulletin Quality

The quality of the final bulletin produced by the *Global Association System* is examined by comparing it to the known events in the synthetic data set. We analyzed the automated bulletin produced by **GA+ESAL** for the 5-day period using a performance analysis tool called **PerfV** that we have used in the past for evaluating **ESAL** performance [*Sereno et al.*, 1993].

The events produced by the *Global Association System* were compared to the known events used to generate the synthetic data after screening those that would not pass the weighted-count and error-ellipse event confirmation tests. Events in the automated bulletin that have the same location solution as a known event in the synthetic data set are classified as *Accepted*. Similarly, events that matched some associations and phase identifications of a known event, but whose location solution is different, are classified as *Modified*. Known events that do not have a corresponding event in the automated bulletin are classified as *Added* (i.e., these are missed by the automated processing). *Rejected* events are those in the automated bulletin that do not correspond to any known event (i.e., false-alarms)[1]. The *Modified* events are subdivided into those that were located <50 km from the corresponding known event, and those that were located >50 km from the known event. Table 4 summarizes the performance of the *Global Association System* for the 5-day synthetic data set in these terms. The results for the swarm day (Day 1) are shown separately from the other four days. The *Added* and *Rejected* events are also subdivided by the number of defining phases.

1. When a known event is split into two or more events, the best match to the known event is classified as *Modified* and the rest are classified as *Rejected*.

Table 4: Analysis of Global Association Bulletin

Class	Day 1 (swarm)	Days 2-5
Known Events	**601**	**1164**
GAassoc+ESAL Events	**621**	**1582**
Accepted Events	*0*	*0*
Modified Events	*448*	*955*
ddist <= 50 km	209	486
ddist > 50 km	239	469
Added Events	*153*	*209*
ndef=3	42	68
ndef=4	35	47
ndef>=5	75	92
Rejected Events	*173*	*627*
ndef=3	66	265
ndef=4	69	229
ndef>=5	38	133

It is not surprising that there are no *Accepted* events because the locations of the known events were not determined from the arrival data. Instead, they were specified by the SDG *Program Package*. Therefore, the automated location solutions would not be the same as the location of the known events even if the automated system produced exactly the same associations and phase identifications.

Approximately 82% of the known events on Days 2-5 were formed by the *Global Association System*, and over half of these had locations within 50 km of the known location. Figure 11 shows a histogram of the distances to *Modified* events which were > 50 km from the known event. Most are within a few degrees, but a number of them are small events with incorrect associations which result in a large error in the estimated location. About 40% of the events in the automated bulletin are false-alarms, but almost 80% of these have ≤ 4 defining phases. A high rate of false-alarms is an unavoidable consequence of pushing the detection threshold down to such low levels (e.g., 3 defining phases). The relatively high proportion of *Added* events (18%) is the major concern. This is surprising for an exhaustive search method like **GAassoc**. Therefore, we investigated some of the factors that can cause events to be missed by the automated processing. It turns out that many of these factors are external to **GAassoc**. For example, random errors in detection features and errors in station processing (**StaPro**) account for many of the *Added* events. Other factors include screening events on the basis of probability

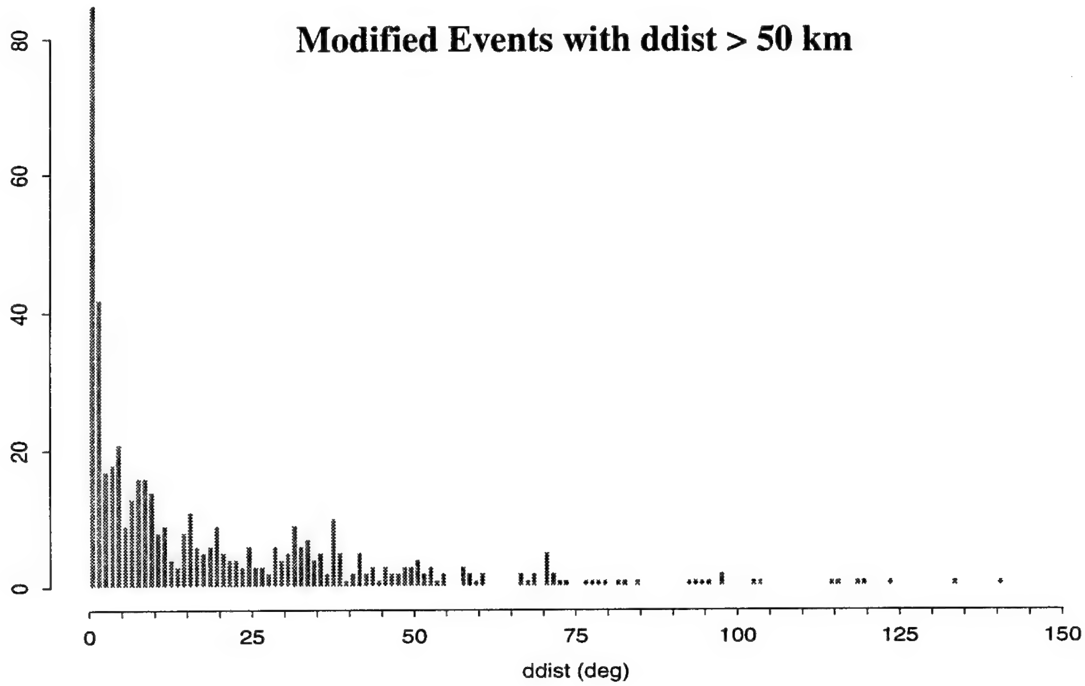


Figure 11. Histogram of Modified Events for Days 2-5 where ddist > 50 km.

of detection and errors in conflict resolution. These factors are explained in more detail below.

The detection attributes in the synthetic data are assumed to follow a normal distribution, and the variance of the distribution depends on signal-to-noise ratio and station type (e.g., see Figure 8). However, these variances were not known to the automated processing system. Instead, nominal values were assumed that did not depend on signal-to-noise ratio. In many cases these nominal values underestimated the true variances. As a result, 11.8% of the phases associated to known events have at least one residual that is more than 3 times the database uncertainty. This prevents association of these phases and could cause marginal events to be missed.

Station processing misidentified the initial wave type of 8.7% of the *P*, *Pn*, *Pg*, and *PKPdf* phases associated with known events. It identified an additional 5.8% of these phases as *P*-type coda phases which means they could not be *DRIVERS*. Most of the station processing errors were for three-component stations for which accurate estimates of slowness are not available. Our experience with IMS data indicates that the identification accuracy (*P* vs. *S*) of the default rules for initial wave-type identification is about 80% for data from three-component stations. The percentage increases to >90% if a trained neural network is used [*Sereno and Patnaik*, 1993]. However, even if neural networks have been trained, some marginal events will be missed because of errors in station processing.

Screening events on the basis of probability of detection is done to remove event hypotheses that are formed by an unlikely combination of stations. This test significantly reduces the rate of false-alarms, but it can also cause real events to missed. The current test depends on the accuracy of the event magnitude estimate, and it could be made more robust by eliminating this

dependency (Section 4.0).

Errors in conflict resolution can also lead to missed events. In these cases, **GAassoc** may have found the known event but it shared associations with other event hypotheses. In the process of resolving conflicts, the correct hypothesis can be lost. The conflict resolution test that we used in this study is very simple and produces some errors. Analysis of the first eight hours of Day 2 of the synthetic data set indicates that this may account for most of the missed events. Twelve events were missed by **GAassoc+ESAL** in that interval, but only 3 of these were missed by **GAassoc**. Improved conflict resolution tests are needed to address this problem.

3.5 Comparison to ESAL

The performance (computational efficiency and bulletin quality) of the *Global Association System* was compared to the performance of **ESAL** using the first eight hours of Day 2 of the synthetic data set. As seen in Figure 7, this interval is representative of the data set during periods of normal seismic activity[1]. **ESAL** used the same time-step interval and the same parameter files as it did in the *Global Association System* except that the *GA* trial origin method was replaced with *Regionals*, *Locals*, *Singles,* and *Doubles* [*Bratt et al.*, 1991, 1994].

3.5.1 Computational Efficiency

Figure 12 summarizes the computation efficiency comparison. The data are grouped into one-hour intervals. There are two bars for each interval. The one on the left is for the *Global Association System* (**GA+ESAL**), and the one on the right is for **ESAL** only. The **GA+ESAL** bar is divided into two parts; the white part shows the CPU time spent in **GAassoc** and the dark part shows the CPU time spent in **ESAL**. This figure shows that the two systems run at very similar speeds for this 8-hour data segment with **GA+ESAL** tending to be slightly faster. However, the **ESAL** component of **GA+ESAL** is taking most of the time (63% on average for this data segment).

Figure 13 shows a breakdown of the CPU time spent in the **ESAL** component of **GA+ESAL**. The initial resolution of **GAassoc** conflicts and the association of secondary phases, the two nominal tasks required of **ESAL**, take a small fraction of the total time. Almost a quarter of the time is spent reading the **GAassoc** preliminary events, a similar amount is spent in other I/O, and an additional 17% is spent in "final" conflict resolution. The main task performed by final conflict resolution is the resolution of conflicts between new **GAassoc** events and *Previous* events that carry over from the preceding time step. This is necessary because **GAassoc** can use phases that are correctly associated with an event in the previous time step to generate new hypotheses in the current time step. Final conflict resolution applies a more conservative approach than initial conflict resolution, and this takes more CPU time. This time could be reduced by optimizing final conflict resolution for use with **GAassoc**.The breakdown in Figure 13 implies that the time taken by the nominal tasks currently required of **ESAL** could be significantly reduced if they were performed within **GAassoc** (thereby eliminating the data transfer between **GAassoc** and **ESAL**). This would also eliminate the time spent in **EServer** and a significant portion of time spent by **GAassoc** writing the preliminary events.

1. The detection density in this interval is approximately 94% of the average detection density for days 2-5 while the event density is approximately 98% of the average event density for days 2-5.

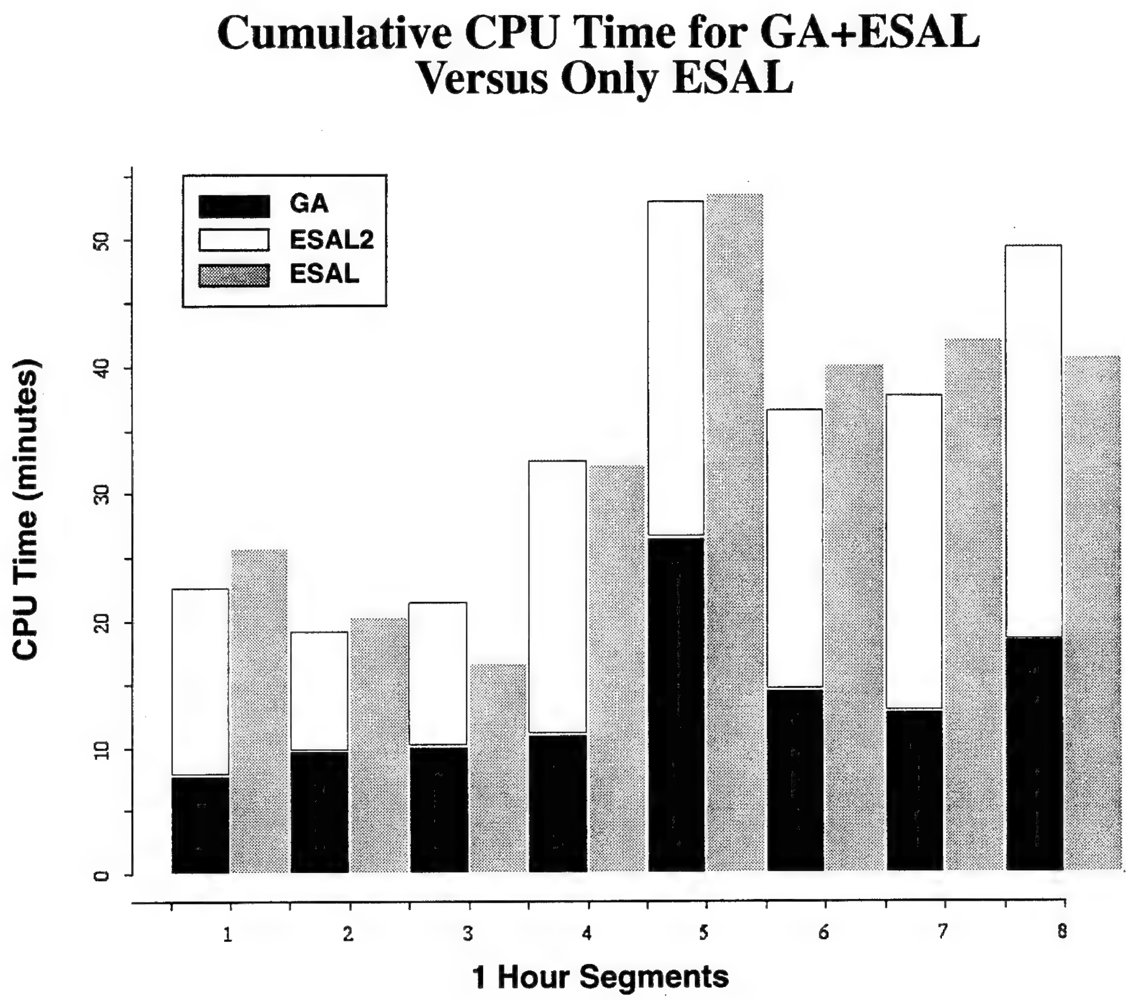


Figure 12. This compares the computational efficiency of the *Global Association System* (**GA+ESAL**) to **ESAL** by itself for eight hours of synthetic data. The figure shows the CPU times for each one hour segment with the breakdown between **GAassoc** and **ESAL** components of the hybrid system. The **ESAL** portion of the hybrid run is labeled **ESAL2** on the figure.

The remaining one third of **ESAL**'s CPU time is spent associating primary phases. This is necessary to repair errors that are made during conflict resolution. This task would be difficult to perform in **GAassoc**. An alternative would be to improve the conflict resolution so that this task is not necessary, but obviously that will increase the CPU time used for that task. These are important issues to be addressed in our follow-on AFTAC Task Order to port the **ESAL** component of the *Global Association System* to **GAassoc**.

The computational efficiency results presented above are expressed in terms of CPU time. The real time taken by **GAassoc** for the eight hours of data was about 1.7 times the CPU time, and the real time taken by **ESAL** was about 1.2 times the CPU time. The workstations used during these tests had varying additional processing loads, and both these ratios are expected to be

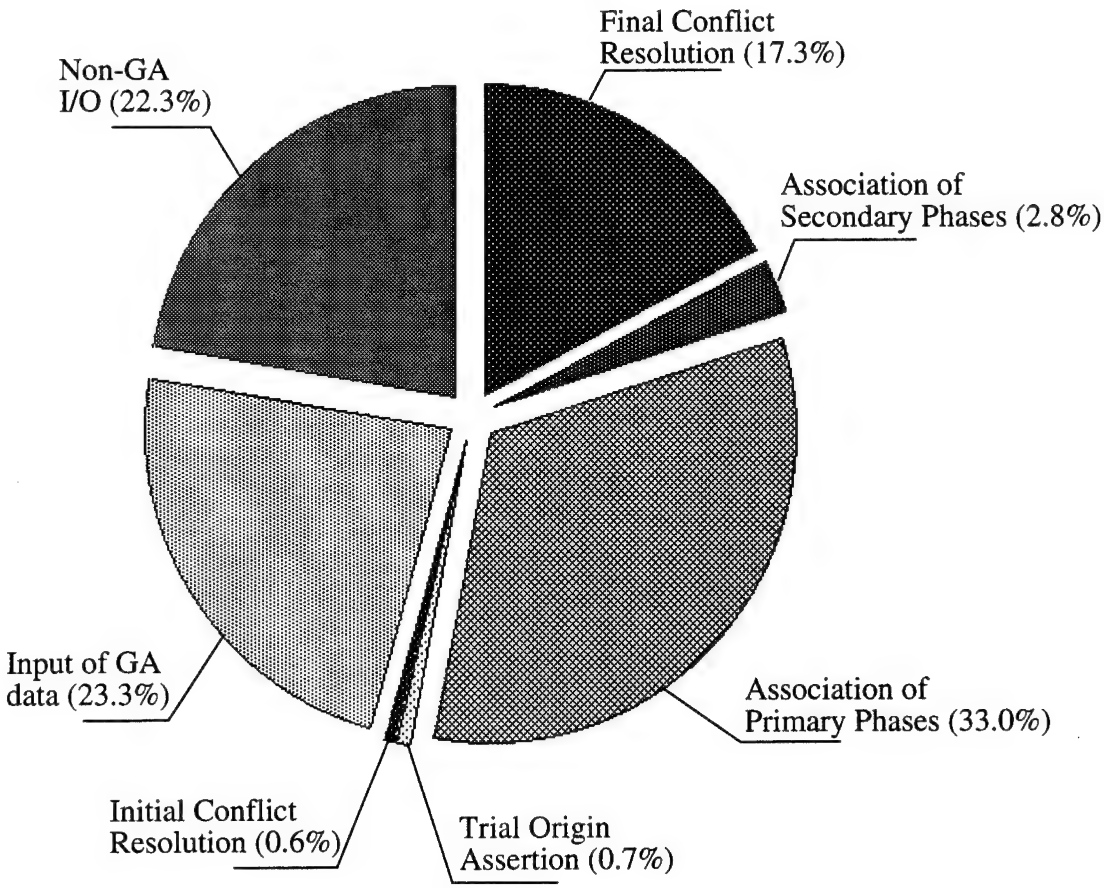


Figure 13. This pie chart shows the percentage of time spent by **ESAL** performing each of its tasks in the *Global Association System* during the first eight hours of Day 2 (a nominal, non-swarm, period).

lower when dedicated workstations are used. In addition, there was time taken by **EServer** that is not shown in Figure 13. It averaged about 2.4 min per 20-min interval.

3.5.2 Bulletin Quality

A comparison of the quality of the bulletins produced by **GA+ESAL** and by **ESAL** alone for this 8-hour period is shown in Table 5, where we have again used **PerfV** to compare the automated bulletins to the known events. The *bulletin quality* is a measure from 0 to 1 computed from:

$$Q = (Pd)^n (1 - Pfa)^{(1 - n)}$$

where *Pd* is a weighted sum of events that were accepted or modified (the weight depends on the distance between the automated event location and the location of the known event), and *Pfa* is the percentage of false alarms, and *n* provides a weighting of these factors which we have set

to 0.75 [*Sereno et al.*, 1993][1]. The **GA+ESAL** System produced 17 fewer false alarms and missed one more event than **ESAL** did when run alone. This shows that the bulletin produced by **GA+ESAL** is similar in quality to the bulletin produced by **ESAL** alone.

Table 5: GA+ESAL and ESAL Bulletins for the first 8 hours of day 2

Class	GA+ESAL	ESAL
Known Events	**92**	**92**
Events in Automated Bulletin	**118**	**136**
Accepted Events	*0*	*0*
Modified Events	*80*	*81*
ddist <= 50 km	36	39
ddist > 50 km	44	42
Added Events	*12*	*11*
Rejected Events	*38*	*55*
Bulletin Quality	**0.71**	**0.69**

3.6 Preliminary Results for the PNDC

The software modules for Station Processing (**StaPro**) and Global Association (**GAcons**, **GAassoc**) were installed at the PNDC at AFTAC. The enhanced version of **ESAL** was not available yet for the PNDC to complete the *Global Association System*. **ESAL** and **EServer** were included in the final software release under this Task Order. A data set was selected for testing and preliminary tuning. Parameters were adjusted, and the results were compared with the event lists obtained from the ADSN Analyst 1 and Analyst 2 database accounts.

The data set included 21 stations from the AFTAC network. The Analyst 1 event list included 17 events. Of these, 2 were removed in the Analyst 2 event list but one of them (from Central Alaska) is assumed to be real. Three events were added in the Analyst 2 event list, resulting in a total of 19 analyst-verified events.

The main objective of the tuning was to maximize the number of real events found by **GAassoc**. A secondary objective was to minimize the number of false-alarms. Station noise levels and signal-to-noise ratio thresholds were adjusted to normalize the probability of detection estimates. Also, station processing parameters were set to the station-specific values currently used in the ADSN **ESAL** configuration. Finally, **GAassoc** user-parameters controlling the event confirmation tests and the number of "first-arrival stations" were adjusted to satisfy the stated objectives.

1. The *bulletin quality* measure is a useful tool for comparing the processing of two automatic association programs run on the same data set but is somewhat difficult to interpret in isolation. This is why we didn't use it to characterize the GA processing of the full five days of data in Table 4.

GAassoc produced 77 event hypotheses for the test PNDC data set. These were grouped into 27 distinct clusters (e.g., see Figure 4). These clusters included 17 of the 19 analyst-verified events. The two events that were missed were located in the Tonga region. Both events were originally formed by **GAassoc**, but then eliminated during event confirmation. One was eliminated because the location error ellipse was too large (the semi-major axis was 1014 km), and the other was eliminated because it did not satisfy the probability of detection test. This latter event was not reported in the Analyst 1 event list.

Two events formed by **GAassoc** were not formed by Analyst 1, but they were added by Analyst 2. Also, **GAassoc** formed two analyst-verified events in Hokkaido, Japan, that were missed by **ESAL**. Some of the events formed by **GAassoc** that were not reported by the analysts are likely to be real (e.g., two are in Alaska). Comparison with independent bulletins would help to establish the validity of such events.

4.0 Recommendations for Future Work

The long-term goal of this effort is to provide a system for automatic association of seismic signals that will handle the large volumes of data in real time that will be recorded by networks designed to monitor compliance with a CTBT. Towards this goal, we recommend the following enhancements to the *Global Association System*:

- *Add Conflict Resolution to* ***GAassoc***

The current system uses **ESAL** to resolve conflicts among event hypotheses produced by **GAassoc**. A significant percentage of the time spent in **ESAL** in the *Global Association System* is in I/O and associating primaries to repair conflict resolution errors. We recommend that the conflict resolution be improved (so that fewer repairs are necessary), and that it be moved to **GAassoc** to eliminate unnecessary I/O.

- *Eliminate the magnitude dependence of the probability of detection screening*

One of the most powerful tools used by the *Global Association System* to screen unlikely event hypotheses is the probability of detection test. However, the current implementation is sensitive to the magnitude estimate. The screening process could be made considerably more robust by eliminating this dependency.

- *Improve screening of false alarms*

The *Global Association System* generates many more preliminary event hypotheses than **ESAL** and consequently tends to produce more false alarms. We recommend the development of additional techniques to supplement the probability of detection test to screen unlikely event hypotheses.

- *Develop a new event refinement module in C*

Event refinement (including the association of secondary phases) is currently performed by **ESAL**. Additional performance improvements could be realized by porting this part of **ESAL** to the C language.

- *Add late-data handling to* ***GAassoc***

The data processing scenario planned for the IDC and PNDC requires the capability to add late-arriving data to incrementally improve a seismic bulletin. For example, data from a secondary network will be used at the IDC to improve the location accuracy of events formed from data in a primary network. **ESAL** has this capability, and we recommend that it be added to **GAassoc**.

- *Develop performance monitoring tools*

We recommend the development of tools to monitor the performance of the *Global Association System* to support knowledge acquisition and system tuning. These tools should monitor performance at each major stage in the processing sequence.

5.0 References

Anderson, J., M. Mortell, B. MacRitchie and H. Turner, Generic Database Interface (GDI) User Manual, *Tech. Rep. SAIC-93/1001*, Science Applications International Corporation, 140 pp., 1994.

Bache, T., S. Bratt, J. Given, T. Schroeder, H. Swanger and J. Wang, The Intelligent Monitoring System Version 2, *Tech. Rep. SAIC-91/1137*, Science Applications International Corporation, 93 pp., 1991.

Bache, T., S. Bratt, H. Swanger, G. Beall and F. Dashiell, Knowledge-based interpretation of seismic data in the Intelligent Monitoring System, *Bull. Seismol. Soc. Am.*, *83*, 1507-1526, 1993.

Bratt, S., G. Beall, H. Swanger, F. Dashiell and T. Bache, A knowledge-based system for automatic interpretation of seismic data to associate signals and locate events, *Tech. Rep. SAIC-91/1281*, Science Applications International Corporation, 73 pp., 1991.

Bratt, S., G. Beall, H. Swanger, F. Dashiell and T. Bache, A knowledge-based system for automatic interpretation of seismic data to associate signals and locate events, *Tech. Rep.* [in progress, revised edition of *SAIC-91/1281*], Science Applications International Corporation, 1994.

Kerr, A. (ed.), Overview GSETT-3, Report prepared by the *GSE Working Group on Planning*, 9 pp., October, 1993.

Le Bras, R., H. Swanger, T. Sereno, G. Beall, R. Jenkins and W. Nagy, Global Association: Design Document and User's Manual, *Tech. Rep. SAIC-94/1142*, Science Applications International Corporation, 67 pp., 1994.

Leonard, S., Automatic global event association and location estimation using a knowledge based approach to generalized beamforming, Proceedings of the 15th Annual PL/ARPA Seismic Research Symposium, *PL-TR-93-2160*, 248-255, 1993.

Ringdal, F. and T. Kværna, A multi-channel processing approach to real time network detection, phase association, and threshold monitoring, *Bull. Seismol. Soc. Am.*, *79*, 780-798, 1989.

Sereno, T. and G. Patnaik, Initial wave-type identification with neural networks and its contribution to automated processing in IMS Version 3.0, *Tech. Rep. SAIC-93/1219*, Science Applications International Corporation, 37 pp., 1993.

Sereno, T., H. Swanger and T. Bache, Progress in Automated Processing and Interpretation of Seismic Data, Proceedings of the 15th Annual PL/ARPA Seismic Research Symposium, *PL-TR-93-2160*, 363-369, 1993.

Taylor, D. and S. Leonard, Generalized beamforming for automatic association, Proceedings of the 14th Annual PL/ARPA Seismic Research Symposium, *PL-TR-92-2210*, 422-428, 1992.

Distribution List

AFTAC/TTR	Technical Report (2 copies)
AFTAC/TTS	Technical Report (1 copy)
CA/STINFO	Technical Report (2 copies)